Impact evaluations, bias, and bias reduction

Eriksen, Steffen

CHAPTER 1

Introduction

1.1 Overview

The cause is hidden, the result is known.[1] These words in Ovid's Methamorphoses (Book IV, 287) may well be the most succinct way to express human fascination and struggles with causal inference. Science as a whole may be defined by the need to organize knowledge around causal explanations and testable predictions. In fact, economics – the science that studies human behaviour under infinite needs and finite resources (Robbins, 1932) – has engaged with the challenge of identifying causal relationships from its inception: Adam Smith's famous "Wealth of Nations" aims from its title to be "an Inquiry into the Nature and Causes" of wealth (Smith, 1776).

Today most economic studies make use of econometric and statistical inference to make causal claims, or assertions that invoke causal relationships between variables (Pearl, 2004)—for example that a certain policy or intervention has a given effect. However, such causal claims may be the target for criticism due to several potential biases in the empirical strategy used (White and Bamberger, 2008). This has given rise to a new generation of studies focusing on experimental designs, also known as randomized control trials (RCTs), which are arguably less affected by these biases than observational studies. Randomized controlled trials randomize the assignment of a certain "treatment" – be it a policy, a medicine, or a simple nudge – and compare outcomes after a certain time with respect to a "control" group. If properly done, it is argued that given a sufficiently large sample, and given that assignment is random, the difference in outcomes measured in RCTs must be attributed to the intervention. However, it is not always possible to randomize. RCTs, for example, can hardly be used to study Marco-interventions in the economy, such as the privatization of healthcare, or to gauge the socio-economic impact of access to credit (except when credit markets are absent ex-ante). Thus, there is a need for other methods that do not rely on randomization to conduct causal inference. In fact, these so-called non-experimental designs still represent a large share of the empirical work in economics (Athey and Imbens, 2017).

This thesis investigates such non-experimental methods in various settings, focusing on how biases can be minimized. It starts at the macro level, considering the impact of national level policy. Next,

[1] Translated from the original Latin: "Causa latet, vis est notissima"

it zooms in at the household level, investigating the impact of microfinance programs on households. Finally, it looks at the individual level, investigating how individual perceptions and behaviour can result in biases.

Although each chapter functions as an independent contribution to the literature, answering its own specific research question(s), they all follow the same idea. Despite many scientists believing that randomization is virtually the only way to (convincingly) establish a causal relationship (Imbens and Wooldridge, 2009), methods not relying on randomization can also (convincingly) establish a causal relationship. The findings in this thesis contribute to the literature on impact evaluation using non-experimental designs. They put bias – and bias reduction – back at the centre of the debate on causal inference, emphasizing the need for continued interest and improvement of non-experimental designs as a fundamental alternative to randomized designs.

1.2 Impact evaluation

What is an impact evaluation and what can it be used for? According to the International Initiative for Impact Evaluation (3ie), a (rigorous) impact evaluation is defined as:

'analyses that measure the net change in outcomes for a particular group of people that can be attributed to a specific program using the best methodology available, feasible and appropriate to the evaluation question that is being investigated and to the specific context' (3ie, 2012).

Following this definition, impact evaluation can help answer key questions about interventions: what works, what does not, where, why and how much? The most important objective of a rigorous impact evaluation is the robust estimation of causal effects that can be attributed to the program, and nothing but the program (Stockmann and Meyer, 2016). This purpose of rigorous impact evaluation puts the question forward about what is meant by 'causal effect'. Following Rubin (1974), the causal effect is defined as the difference in an outcome Y between a unit having been exposed to treatment and the same unit (under the same conditions) having not received the treatment. That is, we are interested in the factual and counterfactual state of a research unit. However, we are not able to observe the same unit in both conditions at the same time. This problem is known as the fundamental evaluation problem (Heckman and Smith, 1995), and a critical difference between a reliable and unreliable impact evaluation is how well the chosen evaluation design measure the counterfactual (Karlan and Goldberg, 2007).

1.2.1 Randomized versus non-randomized

How to determine the counterfactual is the core of evaluation design (Baker, 2000). The methodologies to accomplish this generally fall into two categories, based on how the assignment to the treatment and control group is conducted, randomized (experimental) and non-randomized (non-experimental). Netting out the program impact from the counterfactual conditions can be difficult, as it can be affected by a variety of biases. i.e. results of problems in the evaluation or sampling design that leads to the impact estimate to deviate from its true value (3ie, 2012). RCTs (or experimental designs) are seen as the gold standard for drawing inference about the effect of an intervention (Athey and Imbens, 2017), as they are considered to have the highest degree of internal validity (study design). That is, they are considered the most robust of the evaluation methodologies. The random assignment process, in theory, generates the perfect counterfactual, free from bias; given a large enough sample size. RCTs are a prospective (ex ante) evaluation design as the treatment and control group are selected in advance of the intervention (Karlan and Goldberg, 2007).

Despite its status as the gold standard, there are still several problems associated with running an RCT. First, ethical reasons might render randomization unfeasible. For example, how can it be justified that certain individuals are assigned to treatment while other are excluded from a possible positive treatment (Imbens, 2009). It is possible to address this problem, however, by bringing the control group into the intervention at a later stage. The randomization thus decides when an individual receives the treatment, and not if they receive it. Second, it can be difficult for political reasons to implement an intervention to one group, but not to another. Third, the scope of the intervention might be too broad such that an appropriate counterfactual is not available. This is specially the case when considering macro-interventions such as healthcare privatisations. Fourth, true randomization might be difficult to achieve. In practice, many studies fail to accurately describe their assignment process (Camfield and Duvendack, 2014). It is suspected that pseudo-random methods are often applied for determining the treatment and control group (Goldarce, 2008). Sixth, RCTs, while having a high degree of internal validity, often lack external validity. That is, the generalizability of the results to a larger population (Rothwell, 2005; Lavrakas, 2008). As the ideal conditions required for RCTs virtually never hold (Deaton 2010), outcomes differ both between and within countries. Despite the problems with external validity, many recent papers in economics using a randomized design, do not deal with problems related to external validity (Peters et al., 2016). Finally, while the assumptions required for an RCT to be unbiased are attractive, unbiasedness alone cannot justify the preference for RCTs over other estimators. It might often be desirable to trade in some unbiasedness for greater

precision. That is, is it might be better to have an estimator that is always near to the target, but might be a little of centre, rather than having an estimator that is nearly always wide of the target (Deaton and Cartwright, 2017).

RCTs are not always feasible for various reasons, as outlined above. It is therefore very important to investigate which non-experimental methods can then be used. Non-random methods can then be used as an alternative to randomized designs. Non-random methods aim to generate a control group that resembles the treatment group, at least based on observable characteristics. This is accomplished using econometric methodologies such as matching methods and double difference methods among others (both which will be discussed in later sections). They rely on including control variables to control for differences between the treatment and control group. These designs can be either prospective (like an RCT), where the treatment and control group is selected prior to the intervention, though in a non-random manner, or retrospective, where a control group is identified after the intervention. Identifying the control group after the intervention is for example seen in microfinance, where evaluators may want to evaluate ongoing projects (See chapter 4 of this thesis, or White (2014)).

What makes an impact evaluation expensive is the primary data collection. Non-experimental designs have the advantage that a primary data collection is not always needed if secondary data is available, thus making them a cheaper and faster alternative than their experimental counterpart. Additionally, non-experimental designs are also possible to implement after a program has started. Ethical and political considerations about who should receive the intervention are also less relevant, as the intervention already took place before the impact evaluation started. The primary disadvantage of non-randomized studies lies with the reduced reliability of the results as the methodologies are less robust statistically. That is, they are prone to different statistical biases, which arise when using a non-experimental methodology. It is the objective of these methods to overcome the different types of biases in the best possible way. When done correctly, non-experimental research can make a tremendous contribution to the literature (Reio Jr, 2016).

1.3 Statistical biases

The definition of impact evaluation listed above, states that the context and the questions being investigated are important conditions for what methodology is the best available. What it does not mention is bias. Bias, which is more likely to arise when the best method available is non-

experimental. Statistical biases can arise in many situation e.g. when the sampling process is non-random (self-selection bias), when placement of the intervention is non-random (program placement bias), when observations from the treatment and/or control group drop out during the intervention (attrition bias), or when survey respondents answer in a way that will be viewed favourably by others (social desirability bias). Any of these biases will leave any estimate of the intervention effect invalid, as the model estimating the effect of the intervention would be subject to endogeneity bias.[2,3]

1.3.1 Selection bias

The main challenge for alternative methods relying on observational data is the problem of selection bias (Lensink, 2014). The problem of self-selection arises when individuals tend to select themselves in a certain state, like treated vs not treated (Angrist et al., 1996), given economic or other, usually observed characteristics. In a randomized setting, this problem is solved by generating a control group who were randomly chosen to not participate in the intervention. In a non-randomized setting, the chosen methodology approaches this by attempting to model the selection process to come up with an unbiased estimate of the intervention effect using observational data. The idea is by holding the selection process constant, a comparison between the participants and non-participants can be made. Overall, finding a proper comparison group is difficult. For example, in microfinance, it can be difficult to find a comparison group of non-participants who are similar to the participants. The non-participants should have the same (unobserved) determination, ability and entrepreneurial spirit that lead the participants to join the program in the first place. Impact evaluation that compares participants, who has this determination, ability and entrepreneurial spirit, to non-participants are likely to overestimate the impact of the program. The extent of this over (or under) estimation is then the selection bias which biases the program estimate (Karlan and Goldberg, 2007). Another example relates to macro-economic interventions, where policy makers across different countries decide whether to privatize healthcare or not. Directly comparing countries who went for such a reform to countries that did not, would lead to a biased estimate of the impact of this reform, as the decision to privatize was not random.

[2] That is, our regressor, T_i , representing the effect of the intervention would be correlated with the error term, ϵ_i in the following regression equation: $Y_i = \beta_0 + \beta_1 T_i + \epsilon_i$. Where Y_i is the dependent variable representing some outcome of the intervention.
[3] Each of these sources of endogeneity bias can be shown to a special case of relevant omitted variable bias (Ruud, 2000; Wooldridge, 2002).

1.3.2 Program placement bias

A related issue to self-selection bias is program placement bias. It occurs when an area with the intervention is compared to an area without the intervention. As most interventions are targeted, it is not likely that the two areas would be similar. It is likely that the physical, economic and social environment of the non-participant group would not match that of the participation group in such a case and therefore results in bias. In an RCT, the randomization would have secured a balanced participant and non-participant group in terms of these characteristics, and thus successfully reduces the bias. In observational studies, the bias has to be modelled, similarly to the case with self-selection bias, so that a comparison of the participants and non-participants in the different areas can be made. The problem of program placement bias can be illustrated for example when microfinance institutes choose where to operate. They choose where to operate for a reason. They may target poorer villages, or may start only accepting clients who are better off before they expand to lower their risk. The bias resulting from this can go both ways depending on whether the comparison area is better or worse off than the area where the intervention is taking place.

1.3.3 Attrition bias

Attrition bias is a type of selection bias, caused by a drop out of participants, affecting both the internal validity and the external validity of an evaluation (Jüne and Egger, 2005). Unlike Self-selection bias and program placement bias, an experimental design would not solve for this type of bias. It is thus a bias that is relevant for experimental as well as non-experimental designs. Dropouts from an intervention does not in itself cause bias as long as the dropouts are *random.* If the dropouts among the treatment and comparison group are purely random, then the follow-up survey will still represent the same population as in the baseline survey (Baker, 2000). However, if the dropout pattern is not random, and participants with certain characteristics are more likely to drop out, then attrition bias will be a problem. Dropouts change the composition of the treatment and comparison group, thus influencing the results of the intervention, leading to an over or underestimation of impact of the intervention (Blundell and Costa Dias, 2008). For example, participants in a microfinance program may exit the program prematurely, and the estimation of the impact can be biased in either direction. The direction of the bias depends on the reason for the participants of the microfinance program to drop out. Dropouts, who tend to be worse off than average, would overstate the impact of the

intervention, while dropouts, who tend to better off than average, would understate the impact of the intervention.[4]

One way of managing attrition is by tracking down the dropouts. However, this is rarely done in practice, as it is very costly and time consuming (Duflo et al., 2008). It is more important for the impact evaluation to report the level of attrition, and compare the dropout with participants who remained in the program in terms of observable characteristics to assess whether there are any systematic differences between the two groups.

1.3.4 Social desirability bias

If there is any incentive to lie, survey responds are likely to be biased in whichever direction serves the interest of the respondent (Singer and Ye, 2013). When surveys are applied for collecting data, respondents may not answer some questions truthfully. This is specially the case for questions on sensitive topics. Questions can be seen as sensitive if they are perceived as interfering with private matters, if their raise fear with the respondent about potential repercussions of disclosing the information, or they raise social desirability concerns (Tourangeau and Yan, 2007; Kreuter et al., 2009). Social desirability concerns lays on the thoughts that there are social standards representing a few practices and states of mind and that individuals may distort themselves to seem to follow these standards. Thus, survey respondents may answer sensitive questions in a manner that others would view them favourable to adhere to the underlying social norms (Nederhof, 1986). This can both be over-reporting 'good' behaviour as well as under-reporting undesirable behaviour. There is therefore a discrepancy between the actions of the respondents and their survey response. This discrepancy results in a (social desirability) bias, which like other types of response biases, can have a large impact on the validity of questionnaires and survey, and subsequently the impact evaluation (Furnham, 1986; Nederhof, 1986; van der Mortel, 2008). For example, social desirability bias could play a role in voter turnout reports, where surveys often overestimate voter turnout at elections (Holbrook and Krosnick, 2010). Voting is seen as a democratic duty, and not voting violates this social norm. The bias could also play a role in microfinance, where respondents, self-report their loan use to the microfinance institute (See chapter 4 of this thesis). They are likely to state a social desirable use of their loan proceedings, such that their loan eligibility is not affected negatively. Chapter 5 of this thesis, investigates social desirability bias for the support for Farmers' Market Organizations (FMOs) in

[4] Given the sparse evidence that is available to distinguish between the two types of participants that exits the programs prematurely, the most likely type of participants dropping out are the ones who tend to be worse off on average (Karlan and Goldberg, 2007).

rural Ethiopia. Farmers may feel social pressure to express positive opinions concerning the FMOs, which is not in line with their actions.

1.4 Research objectives

The overreaching objective of this thesis is to study the impact of interventions using non-experimental techniques, studying biases, and how these methods can adequately reduce bias, and serve as a valid alternative to experimental designs.

The chapters individually address the following research questions:
Chapter 2: Do healthcare financing reforms reduce total healthcare expenditures?
Chapter 3: Do microfinance loans improve the wellbeing of its recipients in Bolivia?
Chapter 4: Do microfinance loans improve the wellbeing of its recipients in Ghana?
Chapter 5: Does social pressure and/or opportunistic behaviour influence revealed support for Farmers' Market Organizations?

Each of these chapters consider a different setting, starting at the macro level, and then gradually zooming in to the individual level. The impact evaluation of healthcare financing reforms in chapter 2 uses macro level data from OECD countries. It addresses the self-selection bias that is present when national governments decide to conduct a healthcare financing reform. Chapter 3 then zooms to the household level, evaluating the impact of microfinance loans from a Bolivian microfinance institute. Similarly, chapter 4 uses household level data to assess the impact of a microfinance organization in Ghana. Although chapter 3 and 4 both consider the impact of microfinance programs, the settings are different. Chapter 3 considers a situation where the expansion plans of the microfinance institute can be applied to help drawing causal inference, whereas chapter 4 considers a scenario where the project to be evaluated is already started, and thus the impact evaluation has to be conducted in the absence of a baseline. Chapter 3 and chapter 4, each present a method to reduce the selection bias and program placement bias that is present. In an attempt to explain the results of the impact evaluation, chapter 4 investigates the effect of social desirable behaviour on the reported loan use of the household's microfinance loan. This analyses of social desirable behaviour in chapter 4, sets the connection to chapter 5, where we zoom into the individual level, investigating how bias resulting from social desirable and opportunistic behaviour affects the revealed support for FMOs. The study of social desirability bias in chapter 5 spawns from another impact evaluation, assessing the impact of the

presence of such FMOs, which was part of the joint MFSII evaluation for Ethiopia.[5] Hence, the common theme across all the chapters is bias and bias reduction, when applying non-experimental designs. The next section outlines the (non-experimental) methodologies applied in each of the chapters.

1.5 Methodology

1.5.1 Honourable mentions

In economics, researchers use an array of different strategies when attempting to identify the causal effect of an intervention using observational data. Such identification strategies (Angrist and Kruger, 1999), all try to reduce biases introduced when using non-experimental designs. While there are many more identification strategies available, all having their own merits. This thesis only considers the application of a few of these identification strategies, thus other (very influential) methodologies remain untouched. Some of the methods we do not discuss include two quasi-experimental approaches, namely instrumental variable methods and regression discontinuity designs, and a relatively novel non-experimental approach: synthetic control methods. The literature on instrumental variables is very voluminous, and for reviews on this literature see Imbens (2014), and Chamberlain and Imbens (2004), with the former focusing on the part of the literature concerning heterogeneous treatment effect, and the latter contributing to the literature on weak instruments. The IV approach is not applied in this thesis due to the lack of proper external instruments. The regression discontinuity approach, despite dating back to the 1950s with the work of Thistlewaite Campbell (1960) in the field of psychology, did not enter the economics literature before the turn to the 21st century. The literature has since been reviewed in detail by Imbens and Lemieux (2008), and more recently in Skovron and Titiunik (2015). One of the main conditions for applying a regression discontinuity design, is the existence of a forcing variable. That is, a variable which determines whether an observation belongs to the treatment or control group. The settings discussed in this thesis were not applicable for a regression discontinuity design. The synthetic control method, develop by Abadie et al. (2010, 2014), and Abadie and Gardeazabal (2003), represents one of the most important development in the literature of policy evaluation in the 21st century. Despite its status as a relatively new and interesting methodology, the settings considered in thesis has not been directly applicable for the synthetic

[5] MFS II is the 2011-2015 grant framework of the Dutch Ministry of Foreign Affairs for Dutch NGOs, which is directed at achieving a sustainable reduction in poverty (Joint MFSII Evaluation for Ethiopia, n.d.).

control method to be applied.[6] However, it would be interesting in future work to consider application of this methodology.

1.5.2 Propensity Score Matching

Every econometric evaluation study has to overcome the fundamental evaluation problem and need to address possible existence of selection bias. We are interested in knowing the participants' outcome with and without the treatment. However, we are not able to observe both outcomes for the same participant at the same time. As an approximation, the mean outcome of nonparticipants could be used. This is not advisable, however, as the participants and nonparticipants usually differ even without the treatment (selection bias). The matching approach or more specifically, the propensity score matching approach is one possible solution to the problem of selection. The basic idea is to find nonparticipants who are similar on relevant pre-treatment characteristics. Doing this, differences in outcomes between the participants and this adequately selected group of nonparticipants can be attributed to the intervention. Matching on all relevant characteristics is difficult when the set of covariates is large (curse of dimensionality), and thus Rosenbaum and Rubin (1983b) suggested to use a balancing score, which is a function of all the relevant observable characteristics, thereby reducing the dimensionality, and making matching possible. The propensity score is one of such balancing scores, which is the conditional probability of receiving the treatment given set of observable characteristics. Matching which applies this score is known as propensity score matching (PSM).

Propensity score matching is a widely used method for estimating program impacts (Imbens and Wooldridge, 2009). However, despite the literature on propensity score matching being mature, there are still some interesting application to be considered, as shown in this thesis. PSM is applied in chapter 2 in the context of estimating the impact of healthcare privatizations on a macroeconomic scale. Chapter 3 applies the propensity score as a tool in forecasting the composition of future clients and non-clients in the case of a future expansion of a microfinance program in Bolivia. Chapter 4 applies the propensity score in combination with a double difference methodology to estimate the impact of an ongoing microcredit program in Ghana.

[6] To apply synthetic control method two main identifying conditions have to be fulfilled: First, the treated observation is featured with enough pre- as well as post treatment periods without the treatment. Second, there is an adequate donor pool of observations with the treatment in the complete period from which the synthetic control can be constructed (see. E.g. Kreif et al. 2016).

The key assumption for identification of the PSM estimator is unconfoundedness, introduced by Rosenbaum and Rubin (1983b). This assumption requires that all factors correlated with both the potential outcome, and with the assignment to the treatment are observed. This implies that once controlling for these observable characteristics, the treatment is as good as randomly assigned. Under this assumption, causal interpretation can then be made of the average difference between the group of participants and the group of nonparticipants for the same value of the covariates. Additionally, Imbens (2004) shows that if potential outcomes are independent conditional on the set of covariates, they are also independent of treatment conditional on the propensity score, i.e. the probability of receiving the treatment. That is, all biases due to observable covariates can be removed by conditioning on the propensity score. Estimation of the propensity score can be conducted via estimation of a discrete choice model such as logit or probit model.

An additional assumption when applying PSM is the common support or overlap assumption. The condition implies, that the distribution of the estimated propensity score for the group of nonparticipants completely overlaps the one of the group of participants. It ensures that participants with the same estimated propensity score have a non-zero probability of being both groups. That is, the common support condition makes it certain that any combination of the observable characteristics found in the group of participants can also be found in the group of nonparticipants (Bryson et al., 2002). By restricting the matching to the common support, we avoid comparing the incomparable by dropping a subset of the group of nonparticipants who are not comparable to the group of participants. Combining this with the unconfoundedness assumption, the PSM estimator is the mean difference in outcomes over the identified common support, where observations are weighted by the conditional probability of receiving the treatment.

1.5.3 Difference-in-Difference

In the events that selection characteristics are known and observed they can be controlled for to reduce the bias by utilizing a variety of non-experimental techniques. One such method is propensity score matching as explained above. However, if selection characteristics cannot be observed - be entrepreneurial spirit or motivation in the context of microfinance – then the exclusion of these variables will result in an omitted variable bias in the form of selection bias. If, on the other hand,

these unobserved characteristics are time invariant, then their influence can be removed via a difference-in-difference (or double difference) procedure, and thus reduce selection bias.[7]

Difference-in-difference methods have been an important part of the toolkit for empirical researchers since the early 1990s (Athey and Imbens, 2017). Difference-in-difference methods are typically applied when some groups like villages or geographical areas experience a treatment, such as the introduction of a microcredit loans in their area, while other areas do not. The selection of the treatment and comparison group is not necessarily random, and outcomes are not necessarily the same across the two groups in absence of the treatment. The difference-in-difference estimator produces a credible estimate of the program impact by comparing a treatment and comparison group (first difference) before and after the program (second difference). The underlying assumption of this estimator is that the change in the outcome over time for the comparison group is informative about what the change would have been for the treatment group had the treatment been absent. With this assumption, the average treatment effect can be calculated as the difference between the change in average outcomes over time for the treatment group, minus the change in average outcomes over time for the comparison group.

The thesis considers different application of the difference-in-difference methodology. Chapter 3 of this thesis deviates from the normal difference-in-difference setup by estimating a difference-in-difference model in space rather than in time, following the work of Coleman (1999). His approach builds on the application of a unique survey design, controlling for selection bias by forming a comparison group out of prospective microfinance clients who signed up a year in advance to participate in a village bank program. Chapter 3 extends this methodology, by forecasting potential clients from an area the microfinance institute would expand to in the future. Chapter 4 combines a difference-in-difference estimator with PSM in the sense that the propensity score is applied to define a common support from which a difference-in-difference estimation is conducted, thereby ensuring that the comparison group is similar to the treatment group.

1.5.4 List experiments

Researchers in the field of social science have developed a variety of techniques to obtain truthful responses to sensitive questions. One of such methods is the list experiment (also known as the item

[7] This is also known as the parallel trend assumption. It states that the unobserved heterogeneity does not change over time. Or if it changes over time, then it would change the same for both the treatment and comparison group.

count or unmatched count technique). First introduced by Raghavarao and Federer (1979), it is a technique to increase the number of true answers to sensitive questions through anonymity. The technique yields the proportion of respondents that (dis)agrees with the sensitive item. The technique is simply to apply, can be relatively easily implemented into a larger survey, and is reported by several studies to yield more accurate responses to sensitive questions compared to direct reporting (Holbrook and Krosnick, 2010; Tourangeau and Yan, 2007). The method is implemented by first randomly dividing the group of respondents into two equal sized groups. The first group of respondents then receive a short list of non-sensitive statements, and are asked to count how many (but not which) statements are true for them. The second group of the respond then receives the same list of non-sensitive statements plus one sensitive statement (the item of interest). By then subtracting the mean number of true statements reported in the first group from the mean number of true statements reported in the second group, the proportion of respondents that engages in the sensitive behaviour can be estimated. The proportion of respondents, who admitted to engage in the sensitive behaviour through the list experiment, can then be compared to the proportion of respondents who admitted to engage in the sensitive behaviour via direction questioning. The difference would then tell the proportion of people who are not telling the truth.

The list experiment technique relies on three assumptions (Imai, 2011). First, that the sample of respondents are randomly divided into the two groups. This implies that potential and truthful responses are jointly independent of the treatment variable. Second, the addition of the sensitive statement does not change the sum of affirmative answers to the non-sensitive statements (known as no-design effect). Third, the respond to the sensitive item from the respondents is truthful (no liars). Furthermore, it is assumed that the respondents are not familiar with the mechanism behind the list experiment technique, and therefore do not consciously manipulate their answers. The list experiment approach as described above is applied in Chapter 4 of this thesis to assess the effect of social desirable behaviour on the reported loan use of the household's microfinance loan in Ghana. A shortcoming of the list experiment technique is the lack of ability to control for multiple covariates at the same time. Thus, while being able to conduct the list experiment by subgroups, it was impossible to relate several respondents' characteristics to their answers at the same time. Imai (2011) proposed a multivariate regression technique, solving this problem, which was then applied by Blair and Imai (2012). Chapter 5 of this thesis applies the approach by Blair and Imai (2012) to assess the effect of social desirable and opportunistic behaviour on the revealed support for FMOs in rural Ethiopia.

1.6 Outline

Chapters are organized as follows. Chapter 2 looks at the impact of healthcare financing reforms on total healthcare expenditures for OECD countries. Chapter 3 evaluates the impact of a microfinance program in Bolivia by applying the expansion plans given by the institute to enable causal inference. Chapter 4 considers the evaluation of an ongoing microfinance program in Ghana. Chapter 4 furthermore investigates the effect of social desirable behaviour on the reported loan use of the household's microfinance loan to explain the results of the impact evaluation. Chapter 5 studies the effect of social desirable and opportunistic behaviour on the revealed support for FMOs in rural Ethiopia. Chapter 6 concludes.